Vermaak, Claire

Working Paper

The impact of multiple imputation of coarsened data on estimates on the working poor in South Africa

WIDER Working Paper, No. 2010/86

Provided in Cooperation with:
United Nations University (UNU), World Institute for Development Economics Research (WIDER)

Suggested Citation: Vermaak, Claire (2010) : The impact of multiple imputation of coarsened data on estimates on the working poor in South Africa, WIDER Working Paper, No. 2010/86, ISBN 978-92-9230-324-2, The United Nations University World Institute for Development Economics Research (UNU-WIDER), Helsinki

This Version is available at:
https://hdl.handle.net/10419/54105

UNITED NATIONS UNIVERSITY
UNU-WIDER
World Institute for Development Economics Research

Working Paper No. 2010/86

The Impact of Multiple Imputation of Coarsened Data on Estimates on the Working Poor in South Africa

Claire Vermaak*

July 2010

Abstract

South African household surveys typically contain coarsened earnings data, which consist of a mixture of missing earnings values, point responses and interval-censored responses. This paper uses sequential regression multivariate imputation to impute missing and interval-censored values in the 2000 and 2006 Labour Force Surveys, and compares poverty estimates obtained under several different methods of reconciling coarsened earnings data. Estimates of poverty amongst the employed are found not to be sensitive to the use of the multiple imputation approach, but are sensitive to the treatment of workers reporting zero earnings. Multiple imputing earnings for all workers with missing, interval-censored or reported zero earnings, the proportion of workers earning less than R500 per month falls by almost a third between 2000 and 2006.

Keywords: coarsened data, multiple imputation, poverty, wage distribution, working poor

JEL classification: C81, J31, I32

* University of KwaZulu-Natal, South Africa; E-mail: vermaakc@ukzn.ac.za

This study has been prepared within the UNU-WIDER project on the Frontiers of Poverty Analysis, directed by Tony Shorrocks.

UNU-WIDER gratefully acknowledges the financial contributions to the research programme by the governments of Denmark (Royal Ministry of Foreign Affairs), Finland (Ministry for Foreign Affairs), Sweden (Swedish International Development Cooperation Agency—Sida) and the United Kingdom (Department for International Development—DFID).

ISSN 1798-7237 ISBN 978-92-9230-324-2

The World Institute for Development Economics Research (WIDER) was established by the United Nations University (UNU) as its first research and training centre and started work in Helsinki, Finland in 1985. The Institute undertakes applied research and policy analysis on structural changes affecting the developing and transitional economies, provides a forum for the advocacy of policies leading to robust, equitable and environmentally sustainable growth, and promotes capacity strengthening and training in the field of economic and social policy making. Work is carried out by staff researchers and visiting scholars in Helsinki and through networks of collaborating scholars and institutions around the world.

www.wider.unu.edu *publications@wider.unu.edu*

UNU World Institute for Development Economics Research (UNU-WIDER)
Katajanokanlaituri 6 B, 00160 Helsinki, Finland

Typescript prepared by Janis Vehmaan-Kreula at UNU-WIDER

1 Introduction

Household surveys usually contain earnings data that are coarsened, in that some earnings values are missing through item non-response, while earnings responses consist of both point and interval-censored values. This makes it difficult to construct a continuous earnings variable with which to analyse poverty and inequality. Empirical studies on poverty and inequality in South Africa typically ignore the missing data, and combine point observations with interval midpoints to create a single earnings variable. However, both of these approaches are problematic. By ignoring missing data, researchers implicitly assume that the data are missing completely at random; if they are not, the resulting estimates will be biased. Second, using interval midpoints ignores the distribution of earnings within intervals; any subsequent distribution-based estimates may thus be incorrect.

This paper aims to deal explicitly with the coarsened earnings data by using a multiple imputation technique, and to assess the effect of this technique on estimates of poverty amongst the employed in South Africa. Imputation provides a means of utilizing data that are subject to item non-response, by assigning a plausible value to missing data. In addition, multiple imputation techniques enable the researcher to generate standard errors that properly reflect the uncertainty involved in the imputation process. Using this methodology thus enables the researcher to construct a continuous earnings variable from coarsened data, which can then be used to analyse poverty levels and trends, while simultaneously acknowledging the additional uncertainty arising from the use of imputed data.

There has been considerable research and debate on levels and trends in poverty and inequality in South Africa during the post-apartheid period. The issue of whether poverty and inequality have increased or decreased since the advent of democracy is of great importance, since it goes to the heart of the effectiveness of government's social and economic policies. The emergence of nationally representative household surveys as a data source from 1993 onwards provided researchers with a wide variety of data with which to conduct such studies. Although there is considerable variation in the data sets used, including the Census (Ardington et al. 2006; Leibbrandt et al. 2006), October Household Surveys (Meth and Dias 2004; Leibbrandt et al. 2005), Income and Expenditure Surveys (Leibbrandt et al. 2005; Hoogeveen and Özler 2006), Labour Force Surveys (Meth and Dias 2004; Leibbrandt et al. 2005) and, more recently, the All Media and Products Surveys (van der Berg et al. 2006; van der Berg et al. 2008), most researchers use household income, comprising both earned and unearned income, or household expenditure, as a money metric measure of wellbeing. Most authors focus on the 1995 to 2000 period, and find that inequality rises, but that the direction and extent of any change in poverty is dependent on the poverty line used. For the more recent period, van der Berg et al. (2008) find that poverty has decreased since 2000, while Leibbrandt et al. (2010) find a slight decrease in poverty and an increase in inequality between 2000 and 2009.

There is little doubt that the massive extension of the social grant system since 1994 has moderated South Africa's extremely high levels of poverty and inequality, although the *extent* to which this has occurred remains a matter of considerable debate (Pauw and Mncube 2007; van der Berg et al. 2006; Meth 2006; Leibbrandt et al. 2010). However, it is not yet clear whether labour market developments over the same post-apartheid

period served to reinforce or to counteract the progress made by the social welfare system. The focus of this paper on the working poor thus enables an investigation into the effectiveness of the labour market in providing a living wage, particularly for those at the bottom of the earnings distribution.

In contrast to the wealth of studies using income or expenditure data, relatively few studies focus specifically on the role of earnings in changes in poverty or inequality. Leite et al. (2006) analyse trends in earnings inequality, but not poverty, up to 2004, while Cichello et al. (2001; 2005) analyse earnings dynamics using panel data, but only amongst Africans in KwaZulu-Natal and not focusing specifically on the working poor. Estimates of the number of the working poor at particular points in time are contained in Casale et al. (2004) and in Posel and Casale (2005) as part of a wider study of other issues. They find that the number of the employed who fall below a poverty line of US$2 a day in real terms (2000 prices) more than doubles between 1995 and 2003. The present study therefore investigates more thoroughly how the incidence of poverty amongst the employed has changed since 2000.

The remainder of the paper is structured as follows. The next section briefly reviews the literature on data coarsening, outlining the extent to which it occurs in South African Labour Force Survey data. Section 3 considers the methodology of multiple imputation, how it differs from the usual methods applied to coarsened earnings data in South African household surveys, and how it is applied in this paper. Section 4 presents the data used in the analysis, and compares the estimates of rates of poverty amongst the employed that result from using several different methods of dealing with data coarsening. Section 5 presents trends in poverty and inequality amongst the employed, using the multiple-imputed datasets. Finally, section 6 concludes and explores the implication of these findings for the estimation of poverty in South Africa.

2 The problem of missing and coarsened data

Survey data are frequently incomplete, in that some of the observational units comprising the sample do not respond to one or more of the parts of the questionnaire. When the available (observed) data are analysed as if they make up the complete sample, researchers implicitly ignore the mechanism which created the missing data. In addition to decreased precision that results from analysing a smaller dataset, resulting inferences may also be biased if the observed data differ systematically from the unobserved data.

There are several different ways that non-response for a variable can be generated, as categorized by Rubin (1987). The data are said to be missing completely at random (MCAR) if the missingness depends on neither the observed nor the unobserved (missing) data. The missing data on a particular variable thus constitute a simple random sample of that variable. If the missingness depends on the observed data, but not on the unobserved data, then the data are said to be missing at random (MAR). Under both MCAR and MAR, the missing-data structure is ignorable, since inferences can be drawn on parameters of interest without knowing the nature of the missingness mechanism.

If the missingness depends on both the observed and unobserved data, such that the probability of a value being missing depends on the unobserved value itself, even after conditioning on the observed values, then the data are said to be missing not at random (MNAR). In such cases, the missingness mechanism is non-ignorable, in that it must be taken into account when drawing inferences on parameters of interest.

If the data are MCAR, analysis of the observed data will produce unbiased estimates of parameters of interest, but there will be some loss of precision in accordance with the smaller sample size. However, MCAR is extremely unlikely in practice (Durrant 2005). If the observed data are analysed as if they comprise the complete dataset when the data are MAR or MNAR, resulting parameter estimates may be biased substantially. The extent of the bias is a function of the fraction of missing data (Lacerda et al. 2008: 61).

In addition to data that are entirely missing, data coarsening is also common in surveys. Data are said to be coarsened when they contain some combination of point (actual) responses, interval (bracket) responses and missing values (item non-response). Data on income, assets and earnings in household surveys are often coarsened because survey instruments provide bracketed response options in order to reduce information that would otherwise be lost through item non-response (Heeringa et al. 1997). However, such data are complex for researchers to work with, as it is difficult to combine the different types of data values into a single monetary measure of wellbeing. The mechanism which generates the data coarsening has similar properties to the missingness mechanism; if data are coarsened at random (CAR), then the mechanism which generates the interval censoring and the missing data is ignorable (Heitjan and Rubin 1991). However, unless the data are coarsened completely at random, analysing only the uncoarsened portion of the data will result in biased parameter estimates.

2.1 The extent of data coarsening in the South African data

This study makes use of data collected by the Labour Force Surveys (LFSs) between September 2000 and September 2006. The LFS is a nationally representative household survey, conducted biannually by the national statistics organization, Statistics South Africa (StatsSA), over this period. The LFSs are chosen because of the consistency of the survey instrument in collecting labour market information across time. Thus any trends identified amongst low-earning workers are likely to reflect changes in the labour market, rather than changes in data collection methodology. Using the 2006 survey allows for recent estimates of poverty to be made, while using the 2000 survey allows for a sufficient time period over which to assess trends in poverty. Although the LFSs were conducted biannually over this period, only the September datasets are used, in order to minimize any seasonal variation in earnings. The interim September LFSs are used to assess the consistency of trends across time. An additional advantage of studying the 2000 to 2006 time period is that it encompasses a number of important legislative developments which can be expected to have had an impact on the functioning of the labour market, and hence on poverty amongst the employed. In particular, as a result of the 2002 amendment to the Basic Conditions of Employment Act, minimum wage determinations were extended to a number of sectors in which workers traditionally have been vulnerable, such as domestic work and agricultural wage employment (Department of Labour 2002). Therefore the extent of poverty among the employed, and particularly the wage employed, can be expected to have declined over this time.

The extent to which earnings data are coarsened in the September 2000 and 2006 LFSs is illustrated in Table 1. While most individuals reported earnings as a point figure (that is, a single numerical value), the proportion of workers with such an earnings value falls between the two surveys, and a growing proportion of workers report their earnings as a bracket figure only. A growing proportion of workers report no earnings information. This category includes the responses 'Don't know' and 'Refuse', which are allowed by the questionnaire design. In addition to this data coarsening, a substantial but decreasing proportion of workers report that they have zero earnings, despite working non-zero hours. Therefore, analysing only workers with (positive) point earnings information would mean that other information from 22 per cent of workers in 2000, and 33 per cent of workers in 2006, would be ignored. Even if just workers with zero and missing earnings information are excluded from the analysis, more than ten per cent of the sample of workers is lost. Thus the implementation of methods for dealing with coarsened data, and consideration of how to deal with reported zero earnings, enables a much greater proportion of the data to be analysed than would otherwise be possible.

Table 1: Distribution of earnings values reported by the employed

Proportion of all employed	2000	2006
Point response	0.776	0.667
	(0.006)	(0.011)
Bracket response	0.107	0.231
	(0.005)	(0.010)
Zero earnings	0.079	0.035
	(0.004)	(0.005)
Missing (includes responses 'don't know' and 'refuse')	0.038	0.067
	(0.002)	(0.006)

Source: September 2000 and 2006 LFS.

Notes: (i) standard errors in parentheses; (ii) all estimates are weighted to population levels using weights provided by StatsSA.

3 Imputation methodologies

Imputation is the process by which missing data are filled in using plausible values, so that techniques developed for analysing complete datasets can be used. Single imputation involves replacing each missing value with a single predicted value, to create a single complete dataset. Examples of single imputation methods include mean substitution, regression imputation and hotdeck imputation.[1] However, the fundamental flaw underlying single imputation techniques is that they fail to take into account that imputed values are more uncertain than observed values. Thus the standard errors of any estimates that are subsequently obtained from the singly imputed dataset are likely to be understated, in that they do not reflect this additional uncertainty (Rubin 1987).

Multiple imputation involves applying a stochastic imputation model to the missing data problem. The model is applied *m* times, creating *m* plausible datasets, and thus multiple

1 For a review of imputation techniques, see for example, Durrant (2005) and Lacerda et al. (2008).

imputation produces a distribution of imputed values which reflects the uncertainty involved in the imputation process. Estimates of interest obtained separately from each of the *m* imputed datasets are then combined as follows, using Rubin's rules (Rubin 1987). Let Q_i represent the estimate of interest from the i^{th} imputed dataset, and let U_i represent the variance of that estimate. Then the overall combined point estimate is:

$$\overline{Q} = \sum_{i=1}^{m} Q_i \Big/ m$$

The variance of Q has two components. The average within-imputation variance is given by:

$$U = \sum_{i=1}^{m} U_i \Big/ m$$

The between-imputation component of the variance is:

$$B = \sum_{i=1}^{m} (Q_i - \overline{Q})^2 \Big/ (m-1)$$

The variance of the combined estimate is:

$$T = U + (1 + m^{-1})B$$

The *t* distribution is used for constructing confidence intervals and significance tests,

$$(Q - \overline{Q})T^{-1/2} \sim t_v$$

with degrees of freedom,

$$v = (m-1)\left(1 + \frac{1}{m+1}\frac{U}{B}\right)$$

Thus for large samples, the estimate of $\overline{Q} \pm 1.96\sqrt{T}$ provides a 95 per cent confidence interval for *Q*.

This paper uses a particular multiple imputation technique developed by Raghunathan et al. (2001) for imputing missing values within a complex data structure, when the data are MAR. Called sequential regression multivariate imputation (SRMI), the method can be used to impute both data values that are entirely missing, and those that are known to be located within a particular interval. The method is used not only to impute coarsened earnings data, but also simultaneously imputes missing values of other variables that will be used in later analysis.

The SRMI method proceeds as follows. The variables to be used in the imputation model are ordered from the least to the most amount of missing values. Let the matrix **X** represent all variables that are fully observed, while $Y_1,\ldots,Y_k$ represent the ordered variables that contain missing values. The first imputation begins by regressing Y_1 on **X**, and imputing values for Y_1 using random draws from the appropriate predictive distribution for Y_1. For example, a normal linear regression model is used when Y_i is a

continuous variable, a logistic model when Y_i is binary, and a polytomous logit model when Y_i is categorical. An interval regression model is used to impute values for variables containing both missing and interval values, following a truncated normal distribution when interval values are reported, such that the imputed values are restricted to be within the interval bounds, and a normal distribution without bounds when values are missing.

Since its missing values have now been imputed, Y_1 is appended to the set of predictor variables. Thus Y_2 is now regressed on **X** and the imputed Y_1, and values are imputed for Y_2, and so on until all Y variables have been imputed using all previously imputed variables as covariates. The imputation process is then repeated, updating the regression parameters θ with parameters drawn from the now-complete distribution. This cycle is repeated until the imputed values and parameters converge to a stable distribution. This produces the first imputed dataset. The entire procedure is then repeated *m* times, to produce *m* imputed complete datasets. Estimates of interest, and their standard errors, are produced using Rubin's rules, as outlined above.

3.1 Previous approaches to data coarsening in the South African data

Empirical studies on poverty and inequality in South Africa using household survey data have typically ignored missing earnings or income data, and have assigned each individual or household either the point observation, where observed, or the midpoint of the reported interval, to create a variable with which to measure wellbeing (*cf.* Leibbrandt et al. 2006; Leite et al. 2006). Recently, several authors have examined the nature of interval responses, and the sensitivity of estimates of the earnings distribution to different methods of approximating the distribution within intervals (Posel and Casale 2005; von Fintel 2007).

While providing several alternative methods of dealing with the interval responses, such studies either ignore or fail to deal satisfactorily with missing earnings values. Moreover, none of these studies use multivariate imputation, and thus they fail to account for the additional uncertainty introduced by the imputation process.

Ardington et al. (2006) conduct the first study which multiple imputes missing income values for South African data. They work mainly with the Census 2001 data, and find that income data are missing for 16 per cent of individuals, while a large (but unspecified) proportion of individuals have zero recorded incomes. They therefore apply the SRMI technique to impute income values for these individuals, and then sum individual income across each household and divide by household size, in order to analyse income per capita. They find that SRMI methods produce higher estimates of mean per capita income, and lower estimates of poverty rates, than without using imputation. However, income values in the Census are collected only in brackets. For the majority of their paper, the authors assign each individual the midpoint of their income bracket as their point income. Although they then test the sensitivity of their estimates of poverty and inequality to this approach, they do not do so by applying interval regression SRMI. Rather, they distribute income within each bracket according to the empirical distribution of individual income from the Income and Expenditure Survey conducted in the same year. However, since the IES data are only collected at five-yearly intervals, this technique is not applicable to the LFS data used in the present study. Ardington et al. (2006) find that their estimates are not very sensitive to the

method applied to incomes reported in brackets. Overall, they find that poverty and inequality rise between 1996 and 2001, which confirms the results of other studies which use these datasets without performing multiple imputation.

Two additional studies focus on the methodology of multiple imputation, and its application to South African earnings data. Lacerda et al. (2008) conduct Monte Carlo simulations on LFS data, focusing on the extent to which SRMI can reduce the bias in mean earnings, under MCAR, MAR and MNAR mechanisms. The study recommends that the use of five imputations, with ten iterations, is efficient in reducing bias for a dataset in which 30 per cent of observations are missing. However, this study focuses on imputing missing point earnings observations, rather than imputing interval-censored responses. Daniels (2008) proposes a theoretical approach to dealing with point, interval and missing observations, which includes modelling the ignorability of the coarsening mechanism. However, using this method, missing data is imputed only for earnings, not for other covariates. The study finds that, for September 2000 LFS, the missing data are CAR, but the interval coarsening mechanism is non-ignorable. With respect to poverty rates and Gini coefficients, there is little difference between the non-imputation and multiple imputation estimates, but the midpoint method performs poorly. However, less than three per cent of the earnings sample is missing in the dataset used, and only ten per cent of respondents reported earnings in intervals (Daniels 2008). The effects of using multiple imputation are likely to be greater in samples containing a greater degree of coarsening, such as the September 2006 LFS used in the current study.

3.2 The multiple imputation approach to Labour Force Survey data

The estimates of poverty amongst the employed in this section are presented in two broad categories. In the first category, no multiple imputation is performed. All missing earnings values are excluded, and those who report their earnings in a bracket are assigned the interval midpoints as their estimated earnings. In the second category, a multiple imputation approach is used progressively to produce estimates of interval-censored, missing and reported-zero earnings values.

In order to impute earnings values, SRMI was carried out including standard earnings equation covariates in the imputation model.[2] Thus any missing values for variables such as age, working hours and education were imputed as part of the process of imputing earnings. Of particular interest for this study, the natural logarithm of monthly earnings[3] was imputed using interval regression, in order to deal simultaneously with point observations, interval-censored observations, right-censored observations and missing observations. Thus the imputed log of earnings variable generally consists of the following combination of observations: existing point earnings observations have been retained; interval-censored observations have been imputed to a value within their reported interval, following a truncated normal distribution; and missing observations have been imputed to any value, following a normal distribution. However, in order to

2 Multiple imputation was implemented in Stata using the downloadable function 'ice' (see Royston, 2005), with each of the five multiple-imputed datasets being produced using ten cycles, as recommended by Lacerda et al. (2008). The resulting multiple-imputed datasets were analysed using the downloadable Stata function 'mim'.

3 For the approaches which include earnings values of zero, these values were replaced with a value of one Rand, in order for the logarithm and Gini coefficient to be defined.

test the sensitivity of estimates of poverty and inequality to the imputation method, several different multiple imputed datasets were constructed, by sequentially including each additional type of data coarsening. The parameters of the imputation model, and the distribution of the imputed data, thus differ depending on the approach taken.[4]

Table 2 outlines the approaches within each of the two categories in terms of how interval-censored, missing and zero earnings responses are treated, and the effect of each approach on the sample size, the mean of the natural logarithm of earnings[5] and the Gini coefficient, for the September 2006 LFS. The sample size of the employed, when only above-zero point and interval earnings responses are considered (approach A), is 24,097. Including all workers who report zero earnings (B) increases the sample size to 25,567, and decreases mean earnings. Excluding workers for whom zero earnings are implausible (C) lowers the sample size, and raises mean earnings, slightly. Using SRMI interval regression to impute interval values (D), rather than using interval midpoints (A), makes little difference to mean earnings, but imputing earnings for workers with missing earnings data (E) raises both the mean and the sample size. The sample size reaches its maximum when both zero earnings and imputed values for missing earnings are included. Amongst workers who report above-zero working hours, reports of zero earnings can be treated as true earnings (F), always replaced with missing values to be imputed (G) or imputed on a case-by-case basis according to their plausibility (H). Such differing treatments affect mean earnings substantially, but the full sample size is maintained in each case. Table 2 suggests that the way in which workers who report zero earnings are treated by the study has a much larger effect on summary statistics of the earnings distribution than does the imputation of missing and interval-censored earnings data. Moreover, the Gini coefficient varies more widely amongst different imputation methods than does the mean. This suggests that whether or not to use imputation may be more of a consideration for the treatment of earnings values at the upper than the lower end of the distribution.

The distribution of log real monthly earnings in the 2006 LFS, and the effects of multiple imputation on this distribution, are presented in the kernel density estimates in figures 1 and 2.[6] Without using SRMI, the kernel exhibits ‘bumps’ representing the allocation of the earnings value at the midpoint of the interval to workers who report their earnings in a bracket. For example, the natural logarithm of the midpoint of the R1 – R200 earnings bracket, converted into real terms, is 4.3, which is the location of the first ‘bump’. The main effect of the SRMI for interval and missing data is thus to smooth the kernel, by applying a truncated normal distribution to interval-censored earnings values.

4 For example, the parameters used to produce each of the five multiple-imputed datasets, for the approach in which all missing, interval-reported and reported zero earnings observations are imputed, are displayed in Table A1 in the appendix.

5 Earnings are imputed as a natural logarithm; therefore, the mean is also presented in natural log-form.

6 All density estimates use an Epanechnikov kernel, and the Silverman (1986) rule-of-thumb bandwidth selector.

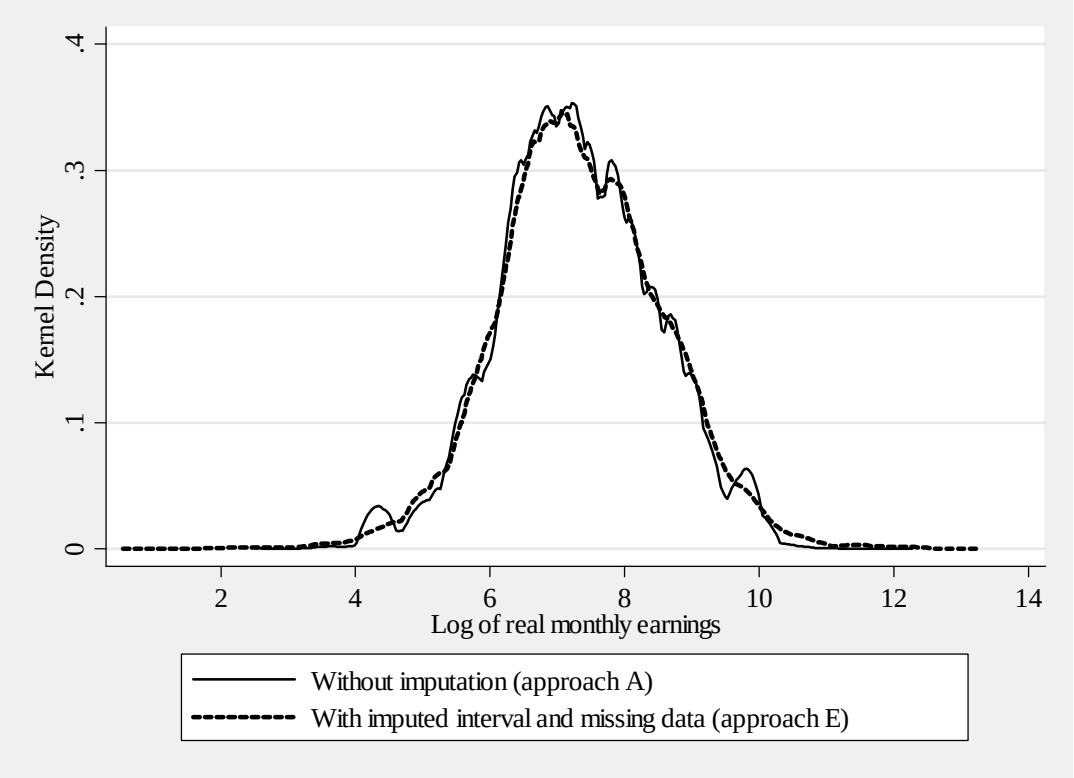
Table 2: Approaches to the treatment of different types of earnings responses

		Treatment of earnings responses			Sample size	Mean of ln(earnings)	Gini
		Interval	Missing	Zero responses			
Approaches without imputation	A	Midpoints	Omitted	Omitted	24,097	7.308 (0.044)	0.585 (0.009)
	B	Midpoints	Omitted	All included	25,567	6.999 (0.076)	0.602 (0.008)
	C	Midpoints	Omitted	Plausible zeroes included; implausible zeroes omitted	25,502	7.012 (0.074)	0.602 (0.009)
Approaches using SRMI	D	Imputed	Omitted	Omitted	24,097	7.308 (0.044)	0.590 (0.009)
	E	Imputed	Imputed	Omitted	25,294	7.347 (0.047)	0.634 (0.015)
	F	Imputed	Imputed	All included	26,764	7.056 (0.077)	0.648 (0.014)
	G	Imputed	Imputed	All imputed	26,764	7.292 (0.049)	0.599 (0.010)
	H	Imputed	Imputed	Plausible zeroes included; implausible zeroes imputed	26,764	7.079 (0.073)	0.606 (0.010)

Source of estimates: September 2006 LFS.

Notes: (i) standard errors in parentheses; (ii) estimates of mean earnings and Gini coefficients are weighted to population levels using weights provided by StatsSA.

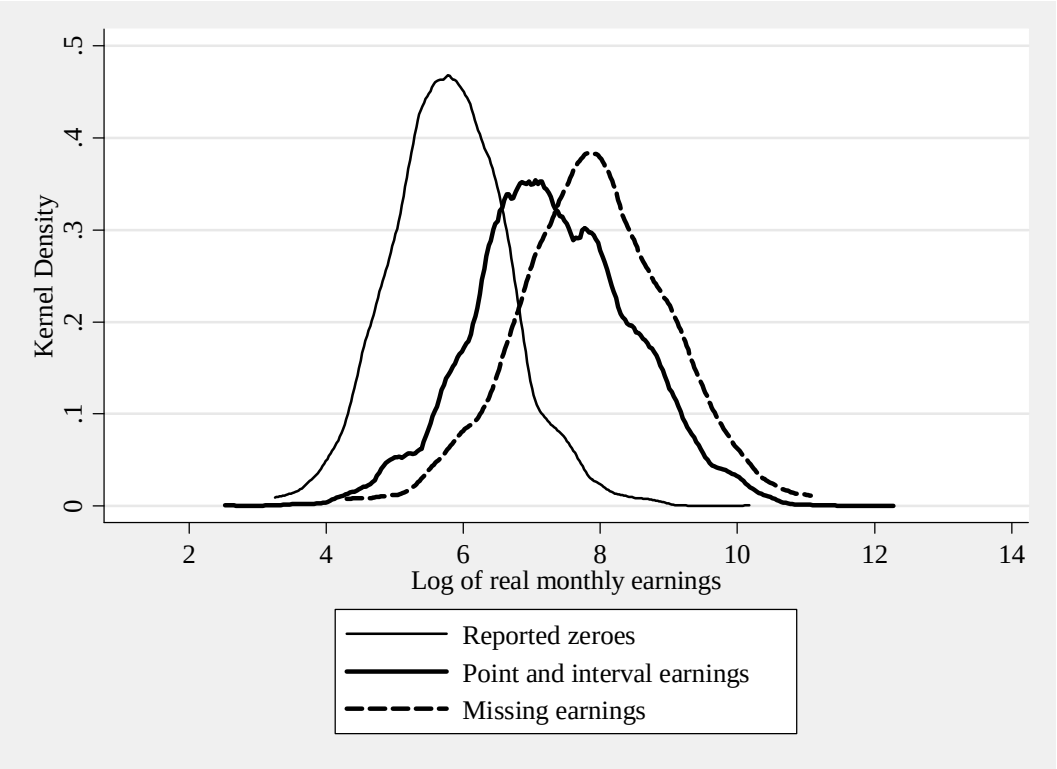
Figure 1: The distribution of earnings, without and with imputation, 2006



Source: September 2006 LFS.

Notes: (i) density estimates are conditional on positive earnings being reported and are weighted to population levels using weights provided by StatsSA; (ii) density estimate for imputed data is shown for the first imputed dataset only.

Figure 2: The distribution of imputed zero, missing and interval-censored earnings, 2006



Source: September 2006 LFS.

Notes: (i) density estimates are weighted to population levels using weights provided by StatsSA; (ii) each sub-sample's density has been estimated separately; (iii) density estimates are shown for the first imputed dataset only.

The application of SRMI produces quite different distributions of earnings for workers who do not report earnings, workers who report zero earnings, and workers who report interval-censored or point earnings. Figure 2 illustrates that imputed values for workers who report zero earnings are substantially lower, and less widely dispersed, than imputed values for other workers, although the imputed values nevertheless lie considerably above zero. Imputed earnings values are highest for workers with missing earnings information, which is consistent with the finding of other authors that workers who do not report their earnings are older, more educated, and more likely to be white and living in an urban area, all of which are characteristics that are associated with higher earnings values (Posel and Casale 2005).

4 The working poor in South Africa

The international literature generally defines the working poor as 'those who work *and* who belong to poor households' (Majid 2001: 2; emphasis in original). However, by identifying poverty at the household level, this definition conflates the earnings of the individual worker with the earned and non-earned income of other members of the household. Changes in the poverty status of an individual worker may then result from changes in his/her individual earnings, changes in the income of other household members, or changes in the composition of the household. Given the substantial increases in social transfers and the changes in household dynamics in South Africa over the study period, the effectiveness of the labour market in redistributing income to the bottom tail of the earnings distribution would be obscured by such a definition.

This study instead defines the working poor as those individuals who work but whose earnings are insufficient to lift them above an individually-defined poverty line. The advantage of such a definition of the working poor is that it enables an analysis of how interactions between the labour market and the characteristics of the individual relate to his/her poverty status at different points in time. The disadvantage of using this definition is that it does not consider income-pooling within the household, and thus can say nothing about how the poverty status of households is affected by changes in the earnings of individual members. Therefore this study in fact amounts to a study of low-earning workers, rather than a general study of poverty.

This study uses two poverty lines, in order to assess the effects of imputation of coarsened data on differently-specified poverty lines, and to assess the extent of changes in poverty at different points in the earnings distribution. The first poverty line is set at R150 per month at real 2000 prices. This poverty line corresponds approximately in 2006 to the boundary between the second (R1 – R200) and third (R201 – R500) earnings brackets in the LFSs, when the brackets are converted into real terms.[7] Although this poverty line has been chosen for its relationship with the earnings bracket, it is close in value to the US$2 per day international poverty line, which amounts to R185 per month in 2000 prices.[8]

The second poverty line is set at R500 per month, at real 2000 prices. This poverty line corresponds to a value slightly below the midpoint of the fourth earnings bracket (R501 – R1000) in 2006, when converted into real terms. This poverty line represents an earnings value approximately 25 per cent higher than the household subsistence level per adult equivalent (Potgieter 1999) in 2000 prices.

4.1 Poverty estimates, without using imputation

The most common method used by researchers to reconcile point and interval earnings data in South African household surveys is to assign interval respondents an earnings value equal to the midpoint of the interval (*cf.* Leibbrandt et al. 2006; Leite et al. 2006). This method is used in this paper for all three of the approaches that do not use multiple imputation. It is not possible to use the empirical intra-band allocation approach of Ardington et al. (2006) since the IES data are only collected at five-yearly intervals, and are thus not compatible with the biannually-collected LFS data.

However, there is a further issue to consider when constructing an earnings variable, in that a substantial proportion of the employed report that they earn zero income. As shown in Table 1, the proportion of all workers who report non-zero working hours but zero earnings is 7.9 per cent in 2000 and 3.5 per cent in 2006. Since the LFS questionnaire asks respondents to report their total salary at their main job, but does not include a prompt for payments in-kind, individuals who do not receive a cash wage, such as those engaged in subsistence agriculture or working without pay in a family business, are likely to report zero earnings. How such workers who report zero earnings

7 Using the average CPI for metropolitan areas for 2006 of 1.34 (Statistics South Africa 2007).

8 Converted from US dollars to South African Rands using purchasing power parity of $1 = R3.90 in 2005 (OECD 2008).

yet have non-zero working hours are treated, and hence whether such zero earnings values are included in the analysis, makes a substantial difference to estimates of poverty.

This study takes three approaches to the treatment of zero earnings. The first (approach A) is to condition the estimates on positive earnings being reported, thus excluding all individuals who report zero earnings. This is the most common approach used by researchers working with the October Household Surveys and Labour Force Surveys. In the second approach (B), all reported zero earnings values are treated as being the genuine earnings of those individuals. In the third approach (C), individuals who report zero earnings are included in the poverty estimates only if it is regarded as 'plausible' that they earn no cash wage. In this approach, earning a wage of zero is considered plausible if, when answering the question 'In the last seven days, did … do any of the following activities, even for only one hour?' (Statistics South Africa 2006), the individual only performed unpaid tasks, such as working in a household business or in subsistence agriculture.[9] More than 95 per cent of workers with zero cash earnings report only performing such types of unpaid work.

Table 3 shows the results of these three approaches. Conditional on positive earnings, 335,000 workers earn less than R150 per month (in real 2000 prices) in 2006, amounting to 2.9 per cent of all workers. 1.8 million individuals, or 15.7 per cent of workers, earn less than R500 per month. These estimates can be regarded as a lower bound for poverty at each poverty line, since they exclude all workers reporting zero earnings. When all such workers are included, the number of the working poor rises by 510,000, resulting in the poverty rate rising to seven per cent at the lower poverty line, and 19.3 per cent at the upper line. These estimates can be regarded as an upper bound for poverty at each poverty line, since they include as poor all workers reporting zero earnings. Since most reports of zero earnings can be regarded as plausible, including only workers with plausible zero earnings produces estimates that are very similar to the upper bound.

Conditional on positive earnings, the Gini coefficient amongst the employed is 0.585. Including either all or some of the reported zero earnings widens the lower end of the earnings distribution, and thus slightly increases the estimate of earnings inequality.[10]

4.2 Multiple imputation of interval and missing earnings values

In this section, sequential regression multiple imputation (SRMI) is used to impute earnings values. Throughout this section, interval regression is used to impute earnings values for the bracket responses, but several different approaches are used for individuals with missing or zero earnings. This allows for comparison with the unimputed estimates. Table 4 presents the estimates of poverty rates, conditional on positive earnings being reported. In the first column of results (A), the estimates without using imputation from Table 3 are repeated, for comparison purposes. In the second

9 Specifically, this includes individuals with zero earnings who gave only responses (d), (e) or (f) to question 2.1 of the LFS questionnaire (Statistics South Africa 2006).

10 As before, reported monthly earnings values of zero were replaced with a value of one Rand, in order for the Gini coefficient to be defined.

Table 3: Poverty amongst the employed estimated without using imputation, by method of treatment of zero earnings

	Approach		
	A	B	C
Poverty Line 1: R150 per month	(positive earnings only)	(incl. all zeroes)	(incl. plausible zeroes)
Working poor ('000s)	335	845	822
	(66)	(190)	(184)
Headcount ratio	0.029	0.070	0.068
	(0.002)	(0.007)	(0.007)
Poverty Line 2: R500 per month			
Working poor ('000s)	1,815	2,325	2,302
	(332)	(455)	(449)
Headcount ratio	0.157	0.193	0.191
	(0.010)	(0.014)	(0.013)
Gini coefficient	0.585	0.602	0.602
	(0.009)	(0.008)	(0.009)

Source: September 2006 LFS.

Notes: (i) poverty lines expressed in real 2000 prices; (ii) standard errors in parentheses; (iii) all estimates are weighted to population levels using weights provided by StatsSA.

Table 4: Poverty amongst the employed, by extent of multiple imputation

	Approach		
	A	D	E
Poverty Line 1: R150 per month	(without imputation; midpoints for intervals)	(imputation for intervals only)	(imputation for intervals and missing data)
Working poor ('000s)	335	335	368
	(66)	(66)	(71)
Headcount ratio	0.029	0.029	0.030
	(0.002)	(0.002)	(0.002)
Poverty Line 2: R500 per month			
Working poor ('000s)	1 815	1 883	2 009
	(332)	(345)	(359)
Headcount ratio	0.157	0.163	0.162
	(0.010)	(0.010)	(0.010)
Gini coefficient	0.585	0.590	0.634
	(0.009)	(0.009)	(0.015)

Source: September 2006 LFS.

Notes: (i) poverty lines expressed in real 2000 prices; (ii) standard errors in parentheses; (iii) all estimates are conditional on positive earnings being reported and are weighted to population levels using weights provided by StatsSA.

column (D), earnings values are imputed for the bracket responses, but not for missing earnings. In the third column (E), earnings values are imputed for both the bracket and missing earnings responses.

Table 4 thus compares the imputation of interval and missing earnings values, to the use of interval midpoints. Since the poverty line of R150 per month corresponds to the boundary between the second and third earnings brackets, imputing values for the intervals (approach D) produces exactly the same estimate of the number and rate of poverty as using the interval midpoints (approach B). However, using the midpoint method assigns the same value to everyone in a bracket, while using interval regression imputation produces a truncated normal distribution within the bracket. Thus the estimate of the depth of poverty would differ by technique, although the poverty headcount does not. There is a difference in estimates between the two techniques at the R500 poverty line, since this line intersects the fourth earnings bracket. The midpoint of this bracket, R560, is greater than the poverty line, thus all individuals reporting earnings in the fourth bracket are classified as non-poor by the midpoint technique. Using imputation, some individuals from within this bracket are classified as poor and others as non-poor. Thus the poverty rate estimated using the imputation technique is slightly higher, at 16.3 per cent, than using the midpoint technique, at 15.7 per cent. The extent to which poverty estimates are affected by using interval regression, rather than bracket midpoints, to impute interval responses thus depends on the extent to which the poverty line bisects an earnings bracket.

Approach E presents poverty estimates when both interval and missing earnings values are imputed. Approximately 33,000 of the 848,000 workers with missing earnings data are classified at the very lower end of the imputed earnings distribution, while a further 93,000 workers earn between R150 and R500 per month. Thus excluding workers with missing earnings data by using the non-imputation approach under-estimates the poverty rate by 0.1 percentage points at the lower poverty line, and by 0.5 percentage points at the upper poverty line.

One of the major contributions of multiple imputation methods, compared to single imputation methods, is that they provide standard errors that properly reflect all sources of uncertainty in the calculation of estimates. Thus although the sample size is larger for the imputation approaches than the non-imputation approaches, the standard errors are also larger, reflecting variability amongst the imputations. Although the approaches that use SRMI produce larger estimates of poverty amongst the employed than the non-imputation methodology, none of the estimates differ by more than one standard error. Thus, conditional on positive earnings being reported, multiple imputation of interval-censored and missing earnings data does not produce significantly different estimates of poverty amongst the employed than the traditional non-imputation methodology.

The imputation of missing earnings data has a much larger effect on the estimated Gini coefficient than on the estimated poverty rate. As was illustrated in Figure 2, most of the distribution of imputed earnings for those with missing earnings information lies substantially to the right of the observed earnings distribution. Including the originally-missing earnings data, in addition to imputed interval-censored earnings, increases the estimated Gini coefficient from 0.590 to 0.634. This increase is significant at the ten per cent level.

In Table 5, estimates of poverty amongst the employed are presented in which SRMI is again used for both interval and missing earnings data. However, the estimates now differ according to the treatment of zero reported earnings. Imputing earnings values for all workers who report zero earnings (approach G), rather than taking all such earnings values at face value (F), roughly halves the number and proportion of workers who earn less than R150 per month. Imputing values only for workers who implausibly report zero earnings results in an estimated 6.3 per cent of workers earning less than R150 per month, and 18.8 per cent earning less than R500 per month. These estimates are quantitatively similar to the figures of 6.8 and 19.1 per cent respectively at the two poverty lines, produced without using imputation (approach C). Once again, none of the SRMI poverty estimates is significantly different to its corresponding non-imputation estimate.

Table 5: Poverty amongst the employed, by method of imputation of reported zero earnings

	Approach		
	F	G	H
Poverty Line 1: R150 per month	(all zeroes included)	(all zeroes imputed)	(implausible zeroes imputed)
Working poor ('000s)	878	421	815
	(195)	(87)	(182)
Headcount ratio	0.068	0.033	0.063
	(0.007)	(0.003)	(0.007)
Poverty Line 2: R500 per month			
Working poor ('000s)	2 519	2 283	2 422
	(482)	(438)	(469)
Headcount ratio	0.195	0.177	0.188
	(0.014)	(0.012)	(0.014)
Gini coefficient	0.648	0.599	0.606
	(0.014)	(0.010)	(0.010)

Source: September 2006 LFS.

Notes: (i) poverty lines expressed in real 2000 prices; (ii) standard errors in parentheses; (iii) all estimates are weighted to population levels using weights provided by StatsSA.

Again, however, the estimates of earnings inequality differ substantially. Including all zero earnings values at face value widens the distribution, raising the Gini coefficient further from approach E. In contrast, the distribution narrows, and estimated earnings inequality declines, when all (G) or some (H) of the zero values are imputed. Compared to approach F, imputing some or all of the reported zero earnings values significantly decreases the Gini coefficient, at the ten and five per cent levels respectively.

Which of the approaches presented above produces the 'right' poverty rate at a given poverty line? It depends largely on what the researcher believes about the validity of zero earnings values, and what they represent. Although individuals working in subsistence agriculture, or without pay in a family business, may indeed receive a zero cash wage, they clearly derive an economic benefit for themselves and their households. In order to reflect this benefit, the remainder of this study will include imputed earnings values for all workers who report zero earnings, provided they report non-zero working hours (approach G). It is important to note, however, that although estimated *levels* of poverty differ when zero earnings are treated according to the various approaches

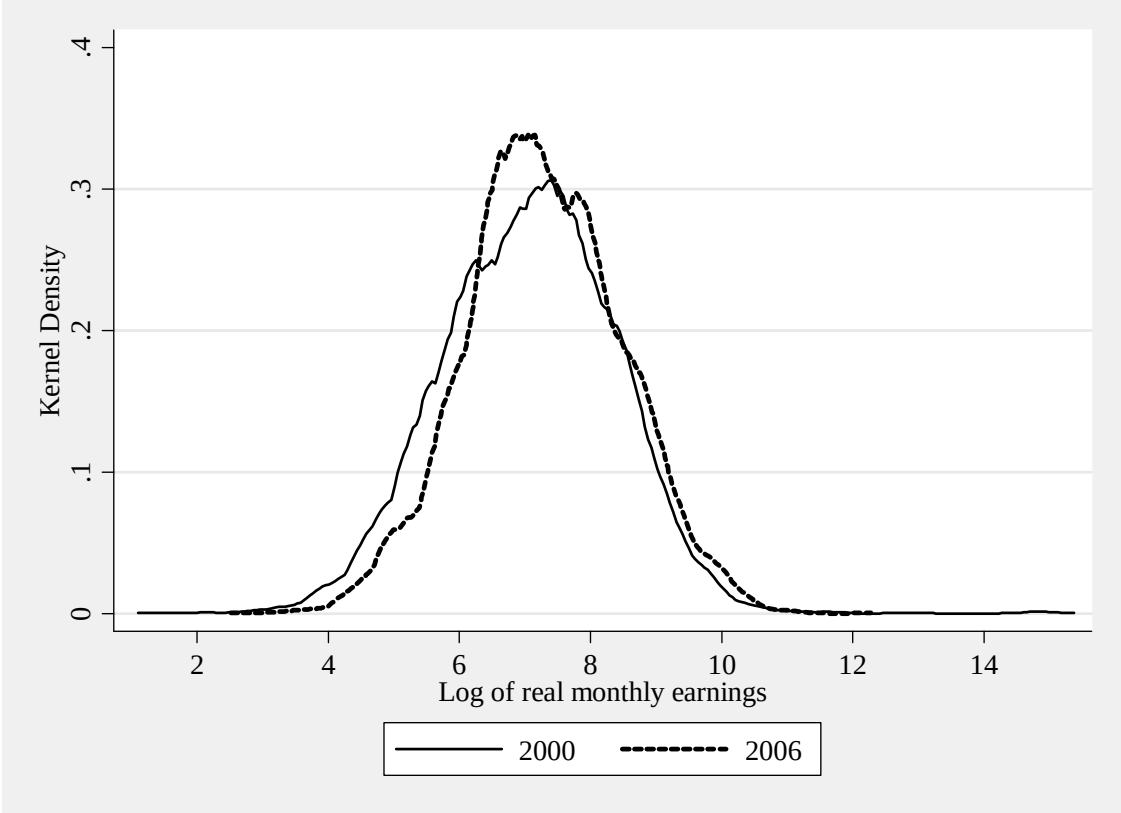
outlined above, both the direction and magnitude of poverty *trends* discussed below is robust to the method of treatment of zero earnings across approaches F, G and H.

5 Trends in poverty amongst the employed

There have been substantial changes in the legislative framework of the South African labour market since the end of apartheid, aimed at setting minimum employment standards, regulating organized bargaining and redressing discrimination. As a result, poverty amongst the employed, and particularly the wage-employed, can be expected to have declined. However, labour market trends over this period may have acted to distribute such gains unevenly amongst the employed. The feminization of the labour force, growing unemployment, informalization of work and growth in self-employment suggest that some types of workers may be crowded into self-employment or jobs in the informal sector which are not covered by the new legislation (Casale 2004; Casale et al. 2004; Bhorat and Cassim 2004).

The kernel density estimate in figure 3 supports these conjectures. There is an unambiguous improvement in real earnings between 2000 and 2006 for those at the bottom of the earnings distribution. However, the log earnings distribution also narrows over time, such that smaller improvements in earnings are achieved higher up in the distribution.

Figure 3: The distribution of real log monthly earnings, 2000 and 2006



Source: September 2000 and 2006 LFS.

Notes: (i) density estimates are weighted to population levels using weights provided by StatsSA; (ii) density estimates are shown for the first imputed dataset for each year; (iii) earnings have been imputed for all missing, interval-reported and reported zero earnings observations (approach G).

Trends in the poverty rates amongst the employed, estimated at the two poverty lines, further illustrate these changes. Table 6 presents poverty and inequality estimates for 2000 and 2006. Approximately 686,000 workers, amounting to 5.6 per cent of the workforce, earnt less than R150 per month in 2000. An additional 2.5 million workers earnt between R150 and R500 per month, such that more than a quarter of all workers earnt less than R500 per month in 2000. However, the rate of poverty amongst the employed declines substantially between 2000 and 2006. More than a quarter of a million workers escape the R150 per month poverty line, while 800,000 fewer are poor as measured by the upper poverty line. The proportion of low-earning workers declines between the two surveys, by more than 40 per cent at the lower of the two poverty lines, and by 30 per cent at the higher poverty line. Thus, on aggregate, workers were significantly better off in 2006 than they were in 2000, but the improvement is larger at the very bottom of the earnings distribution than it is higher up in the distribution.

Table 6: Poverty levels and trends, 2000 and 2006

Poverty Line 1: R150 per month	2000	2006	Change (%)
Working poor ('000s)	686	421	-38.7
	(32)	(87)	
Headcount ratio	0.056	0.033	-42.0
	(0.003)	(0.003)	
Poverty Line 2: R500 per month			
Working poor ('000s)	3 091	2 283	-26.1
	(82)	(438)	
Headcount ratio	0.253	0.177	-30.1
	(0.007)	(0.012)	
Workers earning above R500 ('000s)	9 106	10 599	16.4
	(134)	(1214)	
Gini coefficient	0.786[v]	0.599	-23.8
	(0.046)	(0.010)	

Source: September 2000 and 2006 LFS.

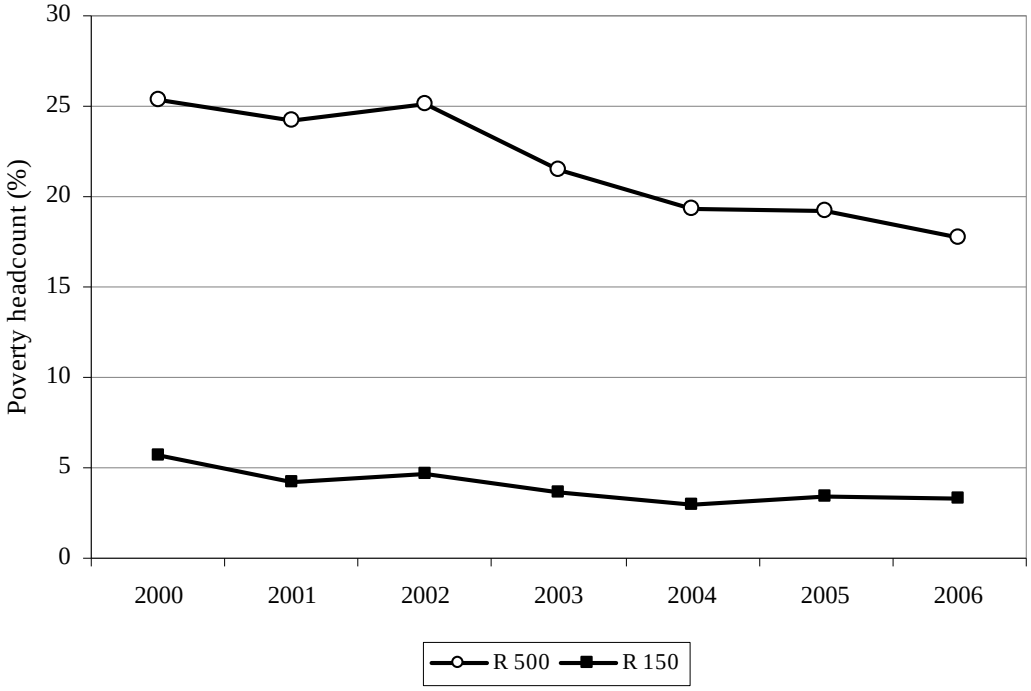
Notes: (i) poverty lines expressed in real 2000 prices; (ii) standard errors in parentheses; (iii) earnings imputed for all missing, interval-reported and reported zero earnings observations (approach G); (iv) all estimates are weighted to population levels using weights provided by StatsSA; (v) including extreme values.

Part of the observed decrease in poverty can be attributed to the decline in the reporting of zero earnings between 2000 and 2006, since a large proportion of the imputed values for workers reporting zero earnings lie below the poverty lines. Conditional on positive earnings being reported, the decrease in the poverty rate is approximately half the size of that reported in Table 6. However, the poverty rate amongst the employed is nonetheless significantly lower in 2006 than it was in 2000.

The decline in working poverty observed in Figure 3 and Table 6 does not appear to be simply the result of low-earning workers losing their jobs, nor an artefact of the chosen endpoints of this study. The number of employed individuals grows over the period, and Table 6 shows that the number of workers earning more than R500 per month increases by 1.5 million, or 16 per cent, over this period. Figure 4 shows that, applying the same SRMI methodology to the interim September LFS datasets, there has been a fairly consistent decline in the proportion of workers who are poor, at both poverty lines, over

the entire 2000 to 2006 period. In particular, the largest decline took place between 2002 and 2004, which is consistent with improvements in labour legislation that occurred over this period.

Figure 4: Trends in poverty amongst the employed, 2000 to 2006



Source: September 2000 to 2006 LFS.

Notes: (i) poverty lines expressed in real 2000 prices; (ii) earnings imputed for all missing, interval-reported and reported zero earnings observations (approach G); (iii) all estimates are weighted to population levels using weights provided by StatsSA.

The Gini coefficient estimated from the imputed 2000 dataset is 0.786, which suggests a very large and significant decline in inequality amongst South African workers between 2000 and 2006. This is surprising, since inequality is known to change slowly over time, and other authors find an increase in total income inequality in South Africa over this period, with only a slight decrease in the labour market contribution to the overall Gini coefficient (Leibbrandt et al. 2010). However, the Gini coefficient estimated here for the September 2000 LFS must be treated with some caution. This particular sample contains a number of implausibly high reported earnings figures. In particular, 12 individuals report monthly earnings in excess of R1 million, despite lacking characteristics commonly associated with high earnings.[11] In contrast, the September 2006 sample contains no individuals with reported real earnings above R300,000. Excluding these 12 individuals from the calculation in 2000 decreases the estimated Gini coefficient for that year to 0.624, suggesting a much smaller decline in earnings inequality over time. While such potential outliers should be more closely examined, rather than simply excluded, especially in the context of a society with extremely high levels of inequality, the magnitude of the reduction in the Gini coefficient demonstrates the sensitivity of inequality estimates to just a few underlying observations, and the

11 None of the 12 individuals is white, and most have not completed secondary education.

difficulty in assessing trends when the extent to which extreme observations are present differs across time.

6 Conclusion

South African household surveys, such as the Labour Force Surveys, contain coarsened earnings data, which consist of a mixture of missing earnings values, point responses and interval responses. The standard approach used by most researchers using these datasets is to create a continuous earnings variable by allocating interval midpoints to bracket respondents, while ignoring missing values. However, such an approach will produce unbiased estimates only if the earnings data are coarsened at random, which is not usually the case.

In contrast, this study uses sequential regression multivariate imputation to produce multiple imputed datasets containing plausible values for both the missing and interval-reported earnings values. Compared to the standard approach, using SRMI significantly raises the estimate of mean earnings in 2006, suggesting that the data were not coarsened at random. However, it does not significantly affect estimates of poverty amongst the employed. Imputed values for missing earnings observations mostly fall above the poverty line, while the imputation of interval-censored responses affects estimates of poverty rates only to the extent that the poverty line bisects an interval.

This study goes on to show that the way in which workers who report earning zero income are treated in the analysis makes a far greater difference to estimates of poverty than does the treatment of missing and interval-reported data. Treating all reported zeroes as genuine, or imputing values only when reported zeroes seem implausible, produces significantly lower estimates of poverty than when earnings are imputed for all reported zeroes.

The study then turns to a brief assessment of trends in poverty amongst the employed between 2000 and 2006. The proportion of workers earning less than R150 per month falls by more than 40 per cent during this time, but the improvement is smaller at a higher poverty line. In particular, the fastest decline in the working poverty rate began in 2002, the year in which the amendment to the Basic Conditions of Employment Act took place.

The analysis of low-earning workers presented here is merely suggestive, and many questions remain to be answered. What sorts of jobs generate such low monthly earnings? Are workers poor because their working hours are insufficient? Are low-earnings workers primary earners, or secondary earners, in their households? Do low-earning workers live in poor households? Thus although a specific focus on earnings is useful, because it enables an analysis of the effects of labour market trends and policies on poverty separate from the effects of the widely-documented extension of the social welfare system, it is also necessary to link low-earning workers with other sources of income in their households, in order to assess overall poverty outcomes.

In conclusion, multiple imputation certainly provides an attractive method of dealing with coarsened survey data. Provided that the imputation model is able to provide a plausible distribution of imputed values, this methodology can reduce non-response bias

while also accounting for the additional variability that arises through imputation. However, implementing multiple imputation imposes costs on the researcher in terms of time and computing resources, both in creating and analysing the multiple imputed datasets. This study has shown that estimates of poverty amongst the employed are not significantly different when implementing SRMI than they are when ignoring missing data and assigning interval midpoints to interval respondents. SRMI does significantly affect estimates of earnings inequality, but such estimates are also extremely sensitive to the presence of outliers at the upper end of the earnings distribution. Thus whether the benefits of the multiple imputation approach outweigh its costs, and whether this methodology becomes standard practice amongst poverty researchers as a result, remains to be seen.

References

Ardington, C., D. Lam, M. Leibbrandt, and M. Welch (2006). 'The Sensitivity to Key Data Imputations of Recent Estimates of Income Poverty and Inequality in South Africa'. *Economic Modelling*, 23: 822–35.

Bhorat, H., and R. Cassim (2006). 'The Challenge of Growth, Employment and Poverty in the South African Economy Since Democracy: An Exploratory Review of Selected Issues'. *Development Southern Africa*, 21 (1): 7–31.

Casale, D. (2004). 'What Has the Feminisation of the Labour Market "Bought" Women in South Africa? Trends in Labour Force Participation, Employment and Earnings, 1995-2001'. Development Policy Research Unit, Working Paper 04/84. Cape Town: DPRU, University of Cape Town.

Casale, D., C. Muller, and D. Posel (2004). 'Two Million Net New Jobs: A Reconsideration of the Rise in Employment in South Africa, 1995-2003'. *South African Journal of Economics*, 72 (5): 978–1002.

Cichello, P. L., G. S. Fields, and M. Leibbrandt (2001). 'Are African Workers Getting Ahead in the New South Africa? Evidence from KwaZulu-Natal, 1993-1998'. *Social Dynamics*, 27 (1): 120–39.

Cichello, P. L., G. S. Fields, and M. Leibbrandt (2005). 'Earnings and Employment Dynamics for Africans in Post-apartheid South Africa: A Panel Study of KwaZulu-Natal'. *Journal of African Economies*, 14 (2): 143–90.

Daniels, R. (2008). 'The Income Distribution with Coarse Data'. Economic Research Southern Africa, Working Paper Number 82. Cape Town: ERSA.

Department of Labour (2002). 'Basic Conditions of Employment Act (No. 75 of 1997) as amended by the Basic Conditions of Employment Amendment Act, 2002'. Pretoria: South African Department of Labour.

Durrant, G. B. (2005). 'Imputation Methods for Handling Item-Nonresponse in the Social Sciences: A Methodological Review'. National Centre for Research Methods Working Paper Series, June. Southampton: National Centre for Research Methods and Southampton Statistical Sciences Research Institute, University of Southampton.

Heeringa, S., R. J. A. Little, and T. Raghunathan (1997). 'Imputation of Multivariate Data on Household Net Worth'. *Proceedings of the Survey Research Methods Section*. American Statistical Association, 135–40.

Heitjan, D. F., and D. B. Rubin (1991). 'Ignorability and Coarse Data'. *The Annals of Statistics*, 19 (4): 2244–53.

Hoogeveen, J. G., and B. Özler (2006). 'Poverty and Inequality in Post-Apartheid South Africa: 1995-2000'. In H. Bhorat and R. Kanbur (eds) *Poverty and Policy in Post-Apartheid South Africa*. Pretoria: HRSC Press.

Lacerda, M., C. Ardington, and M. Leibbrandt (2008). 'Sequential Regression Multiple Imputation for Incomplete Multivariate Data using Markov Chain Monte Carlo'. Southern Africa Labour and Development Research Unit Working Paper Number 13. Cape Town: SALDRU, University of Cape Town.

Leibbrandt, M., J. Levinsohn, and J. McCrary (2005). 'Incomes in South Africa since the Fall of Apartheid'. National Bureau of Economic Research, Working Paper 11384.

Leibbrandt, M., L. Poswell, P. Naidoo, M. Welch, and I. Woolard (2006). 'Measuring Recent Changes in South Africa Inequality and Poverty using 1996 and 2001 Census Data'. In H. Bhorat and R. Kanbur (eds) *Poverty and Policy in Post-Apartheid South Africa*. Pretoria: HRSC Press.

Leibbrandt, M., I. Woolard, A. Finn, and J. Argent (2010). 'Trends in South African Income Distribution and Poverty Since the Fall of Apartheid'. OECD Social, Employment and Migration Working Papers No. 101. Paris: OECD.

Leite, P. G., T. McKinley, and R. G. Osorio (2006). 'The Post-Apartheid Evolution of Earnings Inequality in South Africa, 1995-2004'. United Nations Development Programme, International Poverty Centre, Working Paper 32.

Majid, N. (2001). 'The Size of the Working Poor Population in Developing Countries'. Employment Paper 2001/16. Geneva: International Labour Organization.

Meth, C. (2006). 'What Was the Poverty Headcount in 2004 and How Does it Compare to Recent Estimates by van der Berg et al.?'. Southern Africa Labour and Development Research Unit Working Paper Number 1. Cape Town: SALDRU, University of Cape Town.

Meth, C., and R. Dias (2004). 'Increases in Poverty in South Africa, 1999-2002'. *Development Southern Africa*, 21 (1): 59–85.

OECD (2008). *Purchasing Power Parities and Real Expenditures 2007: 2005 Benchmark Year*. Paris: OECD Publishing.

Pauw, K., and L. Mncube (2007). 'The Impact of Growth and Redistribution on Poverty and Inequality in South Africa'. Development Policy Research Unit, Working Paper 07/126. Cape Town: DPRU, University of Cape Town.

Posel, D., and D. Casale (2005). 'Who Replies in Brackets and What are the Implications for Earnings Estimates? An Analysis of Earnings Data from South Africa'. Economic Research Southern Africa, Working Paper Number 7. Cape Town: ERSA.

Potgieter, J. F. (1999). 'The Household Subsistence Level in the Major Urban Centres of the Republic of South Africa'. Institute for Planning Research Fact Paper No. 107. Port Elizabeth: University of Port Elizabeth.

Raghunathan, T. E., J. M. Lepkowski, J. Van Hoewyk, and P. Solenberger (2001). 'A Multivariate Technique for Multiple Imputing Missing Values Using a Sequence of Regression Models'. *Survey Methodology*, 27 (1): 85–95.

Royston, P. (2005). 'Multiple Imputation of Missing Values: Update'. *Stata Journal*, 5: 188–201.

Rubin, D. B. (1987). *Multiple Imputation for Non-Response in Surveys*. New York: John Wiley and Sons.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Statistics South Africa (2007). 'Consumer Price Index (CPI)'. Statistical Release P0141.1. Pretoria: Statistics South Africa.

Statistics South Africa (various years). 'Labour Force Survey: Unit Records'. Pretoria: Statistics South Africa.

van der Berg, S., R. Burger, R. Burger, M. Louw, and D. Yu (2006). 'Trends in Poverty and Inequality Since the Political Transition'. Development Policy Research Unit, Working Paper 06/104. Cape Town: DPRU, University of Cape Town.

van der Berg, S., M. Louw, and D. Yu (2008). 'Post-transition Poverty Trends Based on an Alternative Data Source'. *South African Journal of Economics*, 76 (1): 58–76.

von Fintel, D. (2007). 'Dealing with Earnings Bracket Responses in Household Surveys – How Sharp are Midpoint Imputations?' *South African Journal of Economics*, 75 (2): 293–312.

Appendix

Table A1: Estimated imputation models for monthly earnings using imputation method G, for LFS September 2006

Dependent variable: natural logarithm of real monthly earnings

	Model 1	Model 2	Model 3	Model 4	Model 5
Age	0.041038	0.041282	0.038526	0.041216	0.042076
	(0.00365)	(0.00394)	(0.00389)	(0.00400)	(0.00414)
Age squared	-0.00039	-0.0004	-0.00037	-0.0004	-0.00041
	(0.00004)	(0.00005)	(0.00005)	(0.00005)	(0.00005)
Male	0.211493	0.211194	0.209807	0.21393	0.210465
	(0.01722)	(0.01725)	(0.01733)	(0.01723)	(0.01738)
African	-0.77415	-0.77999	-0.7748	-0.76662	-0.77763
	(0.02826)	(0.02866)	(0.02832)	(0.02830)	(0.02865)
Coloured	-0.48814	-0.49302	-0.48974	-0.48303	-0.49222
	(0.03324)	(0.03359)	(0.03334)	(0.03327)	(0.03357)
Indian	-0.24548	-0.24885	-0.24653	-0.24157	-0.2497
	(0.05946)	(0.05960)	(0.05949)	(0.05943)	(0.05964)
Cohabiting	-0.15659	-0.15473	-0.15494	-0.15462	-0.15638
	(0.02449)	(0.02452)	(0.02451)	(0.02452)	(0.02455)
Widow/ widower	-0.19092	-0.1926	-0.197	-0.19214	-0.19075
	(0.03847)	(0.03842)	(0.03840)	(0.03832)	(0.03858)
Divorced/ separated	-0.16393	-0.18197	-0.18923	-0.17258	-0.19273
	(0.04290)	(0.04271)	(0.04434)	(0.04242)	(0.04486)
Never married	-0.17714	-0.17836	-0.17897	-0.17636	-0.17875
	(0.01838)	(0.01850)	(0.01848)	(0.01847)	(0.01873)
Primary education	0.165385	0.160954	0.15958	0.159655	0.162135
	(0.02954)	(0.02854)	(0.02855)	(0.02843)	(0.02868)
Incomplete secondary	0.339334	0.335293	0.334644	0.342273	0.337608
	(0.03121)	(0.02992)	(0.02999)	(0.02987)	(0.03021)
Matric	0.692751	0.68597	0.692633	0.695725	0.689398
	(0.03538)	(0.03408)	(0.03413)	(0.03403)	(0.03433)
Post-matric	1.275222	1.268594	1.273565	1.278222	1.268239
	(0.04175)	(0.04065)	(0.04059)	(0.04058)	(0.04092)
Self-reporting	-0.01061	-0.01121	-0.01008	-0.00903	-0.0116
	(0.01542)	(0.01548)	(0.01545)	(0.01542)	(0.01548)
Metropolitan area	0.207862	0.207655	0.208217	0.206601	0.208276
	(0.01665)	(0.01666)	(0.01668)	(0.01665)	(0.01667)
Western Cape	0.225458	0.224581	0.227285	0.229782	0.228639
	(0.03416)	(0.03413)	(0.03412)	(0.03404)	(0.03412)
Eastern Cape	0.015741	0.01608	0.016351	0.01987	0.018437
	(0.03022)	(0.03019)	(0.03020)	(0.03011)	(0.03013)
Northern Cape	0.039367	0.038793	0.039008	0.04356	0.042559
	(0.03258)	(0.03254)	(0.03254)	(0.03246)	(0.03253)
Free State	0.058992	0.058963	0.061029	0.060183	0.060835
	(0.03008)	(0.02998)	(0.03007)	(0.02992)	(0.02996)
KwaZulu-Natal	0.123468	0.122587	0.123225	0.127227	0.125747
	(0.02685)	(0.02679)	(0.02680)	(0.02674)	(0.02678)
North West	0.181282	0.18092	0.183666	0.188398	0.185822
	(0.03318)	(0.03329)	(0.03329)	(0.03313)	(0.03323)
Gauteng	0.249259	0.247757	0.246549	0.249635	0.250043
	(0.02963)	(0.02956)	(0.02958)	(0.02950)	(0.02956)
Mpumalanga	0.195028	0.195602	0.193955	0.196227	0.197084
	(0.02985)	(0.02982)	(0.02977)	(0.02975)	(0.02981)
Household head	0.10754	0.108486	0.111312	0.109863	0.108503
	(0.01684)	(0.01686)	(0.01687)	(0.01687)	(0.01695)
Number of young children in the household	-0.0163	-0.01607	-0.01703	-0.01675	-0.01636
	(0.00879)	(0.00881)	(0.00881)	(0.00879)	(0.00881)

Number of children aged 7 to 14 in the household	-0.02202 (0.00793)	-0.02122 (0.00803)	-0.02037 (0.00802)	-0.0213 (0.00795)	-0.02201 (0.00808)
Hours usually worked per week	0.005665 (0.00056)	0.005594 (0.00056)	0.005678 (0.00056)	0.005598 (0.00056)	0.005651 (0.00056)
Formal sector	0.635718 (0.02952)	0.636031 (0.02985)	0.636276 (0.02939)	0.643449 (0.02966)	0.633946 (0.02973)
Wage-employed	-0.20341 (0.04935)	-0.20198 (0.04942)	-0.20326 (0.04927)	-0.20832 (0.04934)	-0.20234 (0.04934)
Large firm	0.15569 (0.01861)	0.153724 (0.01860)	0.153381 (0.01862)	0.157105 (0.01858)	0.152791 (0.01866)
Agriculture	-0.14196 (0.02719)	-0.14211 (0.02718)	-0.14267 (0.02721)	-0.14213 (0.02718)	-0.14458 (0.02718)
Mining	0.663259 (0.04469)	0.662221 (0.04453)	0.665195 (0.04462)	0.652163 (0.04434)	0.661049 (0.04442)
Manufacturing	0.214889 (0.02443)	0.214991 (0.02438)	0.215853 (0.02444)	0.213973 (0.02441)	0.216309 (0.02441)
Electricity	0.366869 (0.07393)	0.365888 (0.07389)	0.362351 (0.07391)	0.362414 (0.07391)	0.363197 (0.07368)
Construction	0.262191 (0.03283)	0.262863 (0.03268)	0.259798 (0.03288)	0.26223 (0.03269)	0.260309 (0.03272)
Transport	0.304386 (0.03918)	0.305617 (0.03919)	0.302695 (0.03921)	0.303175 (0.03922)	0.305048 (0.03920)
Financial	0.207103 (0.03187)	0.206346 (0.03211)	0.200447 (0.03183)	0.197656 (0.03190)	0.204658 (0.03196)
Community/social services	0.148973 (0.03234)	0.143224 (0.03244)	0.137854 (0.03282)	0.146718 (0.03224)	0.140322 (0.03293)
Private households	0.154382 (0.03382)	0.152454 (0.03445)	0.156115 (0.03369)	0.15985 (0.03421)	0.148398 (0.03471)
Central government	0.443827 (0.04663)	0.447365 (0.04713)	0.4511 (0.04737)	0.442927 (0.04658)	0.451556 (0.04738)
Provincial government	0.487967 (0.03622)	0.492927 (0.03635)	0.49891 (0.03676)	0.486003 (0.03613)	0.496505 (0.03683)
Local government	0.25219 (0.04794)	0.258202 (0.04801)	0.264393 (0.04841)	0.249523 (0.04790)	0.267447 (0.04864)
Government enterprise	0.380797 (0.05669)	0.382117 (0.05676)	0.389686 (0.05682)	0.388127 (0.05679)	0.389127 (0.05659)
Community organisation/ church/ NGO	-0.18878 (0.10846)	-0.19486 (0.11261)	-0.18513 (0.11064)	-0.18921 (0.10936)	-0.19196 (0.11276)
Self-help/ professional association/ union	0.086646 (0.06598)	0.079754 (0.06787)	0.088389 (0.06596)	0.080765 (0.06797)	0.067953 (0.07237)
Own business	-0.0899 (0.05098)	-0.09026 (0.05097)	-0.09059 (0.05096)	-0.09022 (0.05100)	-0.09147 (0.05100)
Intercept	5.643081 (0.10654)	5.652752 (0.11018)	5.694995 (0.10996)	5.628062 (0.10983)	5.634563 (0.11275)

Source: LFS September 2006.

Notes: (i) the sample consists of all of the employed who are aged 15 and older, (ii) all estimates are weighted to population levels using weights provided by StatsSA, (iii) standard errors in parentheses. (iv) The omitted categories for the dummy variables are as follows: for race, 'white'; for marital status, 'married'; for education, 'no schooling', for province of residence, 'Limpopo'; for industry, 'wholesale/retail trade'; and for type of business, 'private business', (v) occupation has been excluded as a covariate due to lack of convergence of the SRMI algorithm when it is included.